A Korean Knowledge Extraction System for Enriching a KBox

Sangha Nam, Eun-kyung Kim, Jiho Kim, Yoosung Jung, Kijong Han, Key-Sun Choi
KAIST / The Republic of Korea
{nam.sangha, kekeeo, hogajiho, wjd1004109, han0ah, kschoi}@kaist.ac.kr

Abstract

The increased demand for structured knowledge has created considerable interest in knowledge extraction from natural language sentences. This study presents a new Korean knowledge extraction system and web interface for enriching KBox, a knowledge base that expands based on the Korean DBpedia. We aim to create an endpoint where knowledge can be extracted and added to KBox anytime and anywhere.

1 Introduction

Information extraction (IE) is an important task in the natural language processing (NLP) field. Various large-scale knowledge bases (KBs) such as Freebase(Bollacker et al., 2008), DBpedia(Auer et al., 2007), and YAGO(Suchanek et al., 2007) are widely used in many NLP tasks. These KBs store knowledge in the form of a triple; for example, (`Les Miserables, author, Victor Hugo`). However, because even large-scale KBs do not contain all the possible knowledge, the knowledge completion task remains crucial in the NLP field. Various approaches can be used for constructing knowledge completion systems, such as knowledge reasoning and extraction. Among them, the task of extracting factual knowledge from unstructured text, such as natural language sentences, is important.

In addition, (Lin et al., 2017) mentioned that certain knowledge is described only in a certain language. For example, the Korean Wikipedia contains much information about Korean culture; similarly, the English Wikipedia contains information about English culture. Moreover, as far as we know, no knowledge extraction system is available for all languages. In addition, building a KB for a specific language requires an ontology schema definition and a knowledge extraction system that is appropriate for that language, as if creating a WordNet (Miller et al., 1990) for each language.

This paper describes a work-in-progress (demo) for building a Korean knowledge extraction system[1] for enriching a KBox[2] knowledge base. The final goal of our research is to build an iterative knowledge learning and extraction system. This web interface plays an important role in accepting new text at anytime and anywhere. Then, knowledge can be extracted from the input text through the web interface and can be accumulated directly in KBox. By doing so, the key modules for knowledge extraction, entity linking and relation extraction (RE), can later learn and improve using this steadily accumulated knowledge. This study makes the following contributions: (1) the first open Korean knowledge extraction system with a web interface and (2) immediately accumulate knowledge that extracted from the proposed system in KBox.

2 System Description

Figure 1 shows the architecture of the proposed demo system. This system has three main parts: Preprocessing, Relation Extraction, and Post-processing. Through the web interface, text is processed se-

[1] http://wisekb.kaist.ac.kr

[2] http://kbox.kaist.ac.kr

Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations, pages 20–24
Santa Fe, New Mexico, USA, August 20-26, 2018.

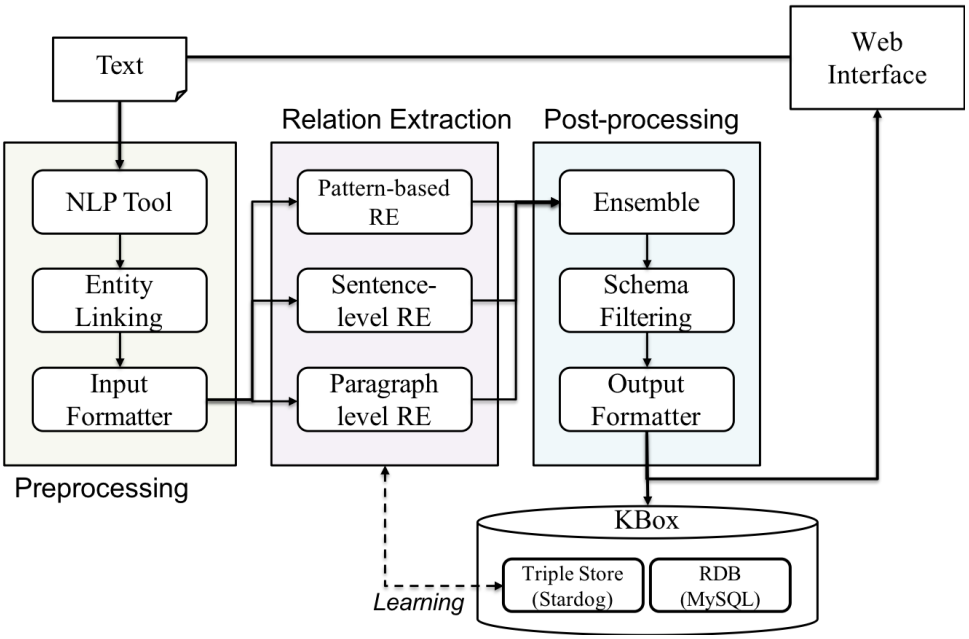

Figure 1: Architecture of the proposed demo system.

quentially by each main parts to extract knowledge and this knowledge is stored in KBox immediately. Details of each part are as follows.

2.1 Preprocessing

Preprocessing involves the following three steps in sequence. **NLP Tool** extracts features such as part-of-speech (POS) tags, dependency parsing, and named entity of input text. **Entity Linking** (Kim and Choi, 2015) links entity mentions in the text with their corresponding entities in KBox. This entity linking system consists of two modules: the entity boundary detection module finds out entity candidates from the text using a bidirectional long-short term memory (LSTM) model with inside-outside-beginning (IBO) and POS tags as features, and the entity disambiguation module takes entity candidates extracted from the entity boundary detection module and selects the most appropriate entity candidate. The system uses a support vector machine (SVM) with entity boundary information and semantic relations between entity candidates, such as entity popularity and inter-entity relations, as features. Korean has an entity made up of single character. Almost all single character entities have different meanings in the same representation, but features that distinguish these different meanings are not enough. Therefore, in our entity linking system, single character entity is not treated as candidate entity mention. **Input Formatter** prepares the input data for each RE model. Because the rule-based RE model use all features generated by the NLP tools, the JSON format was used to effectively deliver this data. Other RE models use the entity-linked text, and a paragraph-level model takes information to distinguish paragraphs using a new-line character.

2.2 Relation Extraction

RE is a task to classify ontological relations between two entities mentioned in a text, and it is a essential for extracting knowledge from natural language sentences. However, even a state-of-the-art RE model (Lin et al., 2017) shows low performance (F-scores 40%–50%). Because it cannot achieve satisfactory performance with just one RE module, we have configured an ensemble with multiple RE models. In the relation extraction step, our system considers not only the entities provided by the entity linking system but also the results of named entity recognition (NER) module as the entity. A new entity that does not exist in KBox cannot be identified by entity linking system, therefore we consider the result of NER as a new entity. Of the many types of NER, only three types of Person, Location, and Organization are considered to be new entities.

The **Pattern-based RE** model (Choi et al., 2016) aims to extract knowledge with high reliability. Human annotators use this model to generate patterns using lexical and syntactic features such as POS, dependency tree, and named entity recognition. This model shows a high precision but low recall, and therefore, scalability is a problem.

Sentence-level RE consists of both convolutional neural network (CNN) (Nam et al., 2018) and LSTM models to address scalability issues and increase recall. These models use distant supervision (Mintz et al., 2009) as a way to collect training data. Distant supervision assumes to collect all the sentences that contain both entities of a triple. Thus, it is widely used as an effective way to automatically create labeled data between a large-scale KB and a corpus. Both CNN and LSTM models use entity-embedded Korean word embedding as input vectors; the CNN model additionally uses vectors for position and POS. The sentence-level RE model is used to reveal the relation between two entities in a sentence; therefore, it is weak at extracting facts that can be found across sentences (paragraph).

One of the differences between Korean and English is the zero anaphora. In Korean, repeated subjects are frequently omitted in the latter sentence. To address this problem, the **Paragraph-level RE** model (Kim and Choi, 2018), which is useful for estimating omitted subjects and predicting relations, explores the incorporation of global contexts derived from paragraph-into-sentence embedding as a means of compensating for the shortage of training data in distantly supervised RE. This model specifically performs zero subject resolution through entity-relation-based graph analysis to find a central entity. The central entities are selected from each paragraph by calculating the out-degree centrality based on the network model of the entity graph using the knowledge base triples. This allows us to learn RE models for informal sentences and has the advantage of compensating for a shortage of training data in the DS-based approach to null subject languages.

2.3 Post-processing

Rather than independently determining the knowledge extracted from each RE module, it is important to combine the results of all modules. We have created an **Ensemble** module based on two concepts: (1) knowledge with high score from one module and (2) same knowledge extracted in multiple modules. A KB should be built based on an ontology schema. Unfortunately, automatically extracted knowledge includes some errors, many of which do not fit the ontology schema. **Schema Filtering** identifies invalid triples and filters them out using domain and range definition of each relation(property) based on two concepts: (1) If the domain or range of relation and subject or object entity types do not match, the triple will be filtered. (2) If the type of entity is not defined, the triple will pass with low calibrated score. **Output Formatter** produces two types of output data. The first is the JSON format for the web interface, and the second is the tab-separated values format that includes triple, triple score, source module, and source sentence for adding to KBox. Through this series of processes, knowledge is extracted and accumulate continuously in KBox.

2.4 KBox

KBox is a new KB that expands Korean DBpedia[3]. KBox consists of two types of storage: One keeps track of both candidate and reliable triples by MySQL, and the other stores only the reliable triples in the former storage by Stardog, a type of a triple store. All the information about all the triples, such as triple scores, the source module, and the source sentence, are stored by MySQL. The reliable triples consist of (1) the initial triples extracted from the Wikipedia infobox (DBpedia) and (2) the automatically extracted triples using the proposed system with a score above 0.9. The expansion of KBOX in Korean DBpedia is three-fold. First, the class hierarchy follows that used in DBpedia[4], but property definitions are revised and strengthened. The domain and range of each property are defined to be common to each language; however, we examined the triples in the Korean DBpedia and found that the schema can be defined more precisely or need to be modified. We then revised the KBox schema by performing instance-based domain range inference. Second, KBox has improved on the triple compared to Korean

[3]http://ko.dbpedia.org/

[4]http://mappings.dbpedia.org/server/ontology/classes/

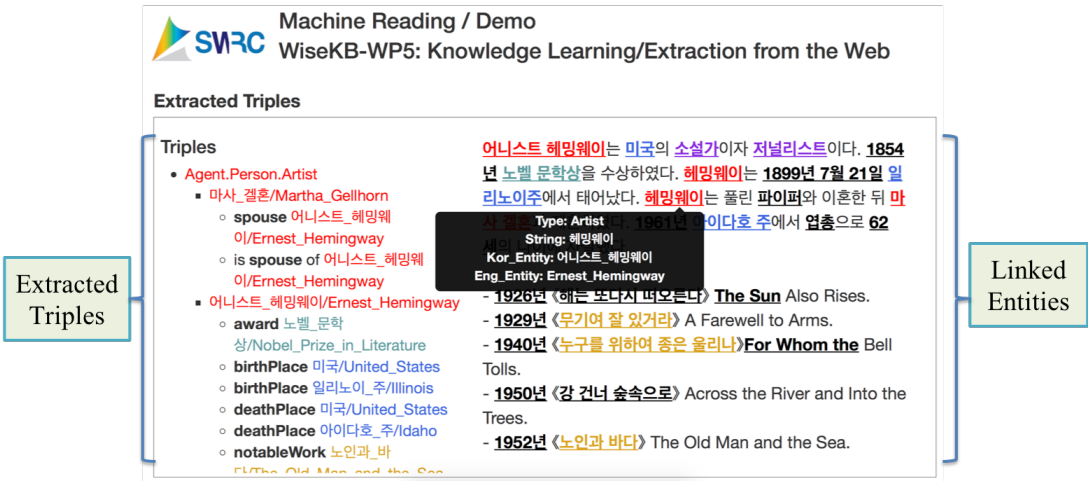

Figure 2: Screenshot of an extracted knowledge from a sample input

DBpedia. First, we defined 763,974 types for 81,991 entities based on the `sameAs` link information of the English DBpedia and Wikidata more than Korean DBpedia. Second, we converted local properties into ontological properties using a mapping table which was manually created by three expert annotators. As a result, 1,678,163 triples represented by Korean local properties were converted, for example, `prop-ko:chul-saeng-ji` to `dbo:birthPlace`. This makes it possible to express a triple represented by a different relation name for the same knowledge in one unified relation. Third, automatically extracted triples are added from this proposed demo and other batch processes in our own server.

3 Demonstration

Figure 2 shows a screenshot of the proposed demo system. Our demo system basically uses Korean natural language sentences as an input. The extracted knowledge is presented to the user in two forms. First, the entity linking results are displayed in color and underline on the input text. When you move the mouse over an entity, the entity type, lexical mention, and Korean and English entity names are displayed. Second, the triples are displayed sorted and rolled up by entity. To demonstrate effectively to users who do not use Korean as a native language, English entities corresponding to Korean entities are displayed together. The source code for our demonstration system has been released[5] under a CC BY-NC-SA license.

4 Conclusion

This study develops a new Korean knowledge extraction system for enriching a KBox. The main contribution is to improve the user accessibility through a web interface, and to provide a Korean knowledge extraction system. Furthermore, new knowledge extracted from the web interface is continuously accumulated in KBox. The core knowledge extraction core modules such as entity linking and RE have laid the foundation for improving the learning model based on the enhanced KBox.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

[5]https://github.com/machinereading/wisekb-demo

GyuHyeon Choi, Sangha Nam, Dongho Choi, and Key-Sun Choi. 2016. Filling a knowledge graph with a crowd. In *Proceedings of the Open Knowledge Base and Question Answering Workshop (OKBQA 2016)*, pages 67–71.

Youngsik Kim and Key-Sun Choi. 2015. Entity linking korean text: An unsupervised learning approach using semantic relations. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 132–141.

Eun-kyung Kim and Key-Sun Choi. 2018. Incorporating global contexts into sentence embedding for relational extraction at the paragraph level with distant supervision. In *LREC*.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 34–43.

George A Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J Miller. 1990. Introduction to wordnet: An on-line lexical database. *International journal of lexicography*, 3(4):235–244.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Sangha Nam, Kijong Han, Eun-kyung Kim, and Key-Sun Choi. 2018. Distant supervision for relation extraction with multi-sense word embedding. *Global Wordnet Conference, Workshops on Wordnets and Word Embeddings*.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.